Universitat Oberta de Catalunya

Patricia Lázaro Tello
plazarotello@uoc.edu

A3: Modelización predictiva

Patricia Lázaro Tello

Índice general

Patricia Lázaro Tello
plazarotello@uoc.edu

Carga de datos

En primer lugar, se procede a cargar el fichero de datos `dat_Air.csv` y comprobar sus variables.

```
csv <- 'dat_Air.csv'
air <- read.csv(csv)

air.dim <- dim(air)

head(air, n=3)
```

```
##   Estacion latitud longitud     Fecha Periodo SO2  H2S NO NO2 NOX O3 PM10 PM25
## 1       12      44     -5.7 11/7/2021      24   1 18.3  5  36  43  5   95   30
## 2       12      44     -5.7 11/7/2021      23   1  6.2  2  22  24 10   80   23
## 3       12      44     -5.7 11/7/2021      22   1  4.3  1  19  20 14   38   15
##   BEN TOL MXIL Dir_Aire  Vel Tmp HR  PRB RS LL
## 1 1.6 2.3  3.2      260 1.49  12 93 1025 36  0
## 2 0.7 1.8  2.7      201 0.98  12 93 1025 36  0
## 3 0.6 1.1  3.1      296 0.98  11 92 1025 37  0
```

```
colSums(is.na(air))
```

```
## Estacion  latitud longitud    Fecha  Periodo      SO2      H2S       NO
##        0        0         0        0        0       54       50       63
##      NO2      NOX        O3     PM10     PM25      BEN      TOL     MXIL
##       63       52        62       43      142       88       86       86
## Dir_Aire      Vel      Tmp       HR      PRB       RS       LL
##        0        0        0        0        0        0        0
```

El fichero contiene 7.464 muestras con 23 atributos cada una. Se observan varias **variables con valores nulos**, que se proceden a eliminar:

```
air <- na.omit(air)
```

Patricia Lázaro Tello
plazarotello@uoc.edu

1 Regresión lineal

1.1 Estudio de correlación lineal

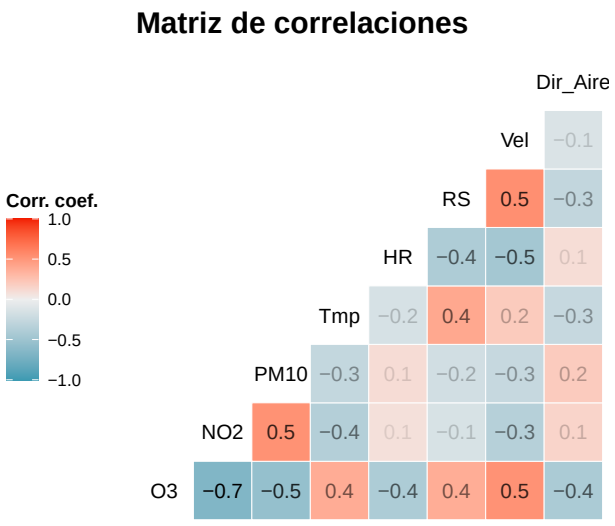
Se procede a comprobar la existencia de correlaciones entre los contaminantes O_3, NO_2, PM_{10} y las variables meteorológicas Tmp, HR, RS, Vel y Dir_Aire.

```
air.corr.vars <- air %>% dplyr::select(O3, NO2, PM10,
                                       Tmp, HR, RS, Vel, Dir_Aire)

cor(air.corr.vars, method = 'pearson', use = 'complete.obs')
```

```
##              O3    NO2  PM10   Tmp     HR    RS   Vel Dir_Aire
## O3         1.00 -0.661 -0.55  0.38 -0.403  0.37  0.52   -0.416
## NO2       -0.66  1.000  0.52 -0.36  0.082 -0.14 -0.35    0.150
## PM10      -0.55  0.523  1.00 -0.27  0.105 -0.18 -0.25    0.202
## Tmp        0.38 -0.362 -0.27  1.00 -0.152  0.40  0.20   -0.252
## HR        -0.40  0.082  0.11 -0.15  1.000 -0.40 -0.46    0.099
## RS         0.37 -0.142 -0.18  0.40 -0.400  1.00  0.54   -0.276
## Vel        0.52 -0.346 -0.25  0.20 -0.459  0.54  1.00   -0.124
## Dir_Aire -0.42  0.150  0.20 -0.25  0.099 -0.28 -0.12    1.000
```

```
ggcorr(air.corr.vars, method=c('complete.obs', 'pearson'), name='Corr. coef.',
       label=TRUE, label_alpha=TRUE, legend.position='left') +
  title.centered + ggtitle('Matriz de correlaciones') +
  theme(legend.title=element_text(face='bold', size=10))
```



Analizando los resultados de la matriz de correlaciones anterior, se observa que **la**

Patricia Lázaro Tello
plazarotello@uoc.edu

cantidad de O_3 en el aire influye en las variables meteorológicas:

Cuanto más O_3 hay en el aire, existe mayor temperatura ($corr(O_3, Tmp) = 0.4$), mayor radiación solar ($corr(O_3, RS) = 0.4$) y mayor velocidad lleva el viento ($corr(O_3, Vel) = 0.5$).

También existe menos humedad en el ambiente ($corr(O_3, HR) = -0.4$) y más al noreste (NE) sopla el viento ($corr(O_3, Dir\_Aire) = -0.4$).

Respecto a **los otros contaminantes (NO_2 y PM_{10}), afectan menos a las variables meteorológicas**:

Se observa cierta correlación negativa entre NO_2 y Tmp ($corr(NO_2, Tmp) = -0.4$), y entre NO_2 y Vel ($corr(NO_2, Vel) = -0.3$). Esto quiere decir que cuanto más NO_2 hay en el ambiente, más frío hace y menos sopla el viento.

Se observa una pequeña correlación negativa entre PM_{10} y Tmp ($corr(PM_{10}, Tmp) = -0.3$), y entre PM_{10} y Vel ($corr(PM_{10}, Vel) = -0.3$); cuantas más partículas en suspensión hay en el aire menos temperatura hace y menos sopla el viento.

También es interesante analizar las correlaciones entre contaminantes y las que se dan entre variables meteorológicas:

Existe **correlación negativa entre O_3 y PM_{10}** ($corr(O_3, PM_{10}) = -0.5$), y una **correlación negativa fuerte entre O_3 y NO_2** ($corr(O_3, NO_2) = -0.7$). Cuanto más O_3 haya en la atmósfera, menos NO_2 y partículas en suspensión habrá.

También existe una **correlación positiva entre NO_2 y PM_{10}** ($corr(NO_2, PM_{10}) = 0.5$); cuanto más NO_2 haya en el ambiente, más partículas en suspensión habrá.

Cuanta más humedad haya en el ambiente, menos radiación solar ($corr(HR, RS) = -0.4$) y menos soplará el viento ($corr(HR, Vel) = -0.5$). A su vez, cuanta más radiación solar hay, más fuerte sopla el viento ($corr(RS, Vel) = 0.5$).

De estos datos se puede concluir que **el contaminante que más afecta a las variables meteorológicas estudiadas es el O_3**, produciendo un aumento de la temperatura, la radiación solar y la velocidad del viento, y un descenso de la humedad relativa. Además, cuanto más O_3 hay concentrado en el ambiente, más al noreste sopla el viento.

La concentración de NO_2 y partículas en suspensión afecta menos, produciendo un descenso de la temperatura y de la velocidad del viento.

Patricia Lázaro Tello
plazarotello@uoc.edu

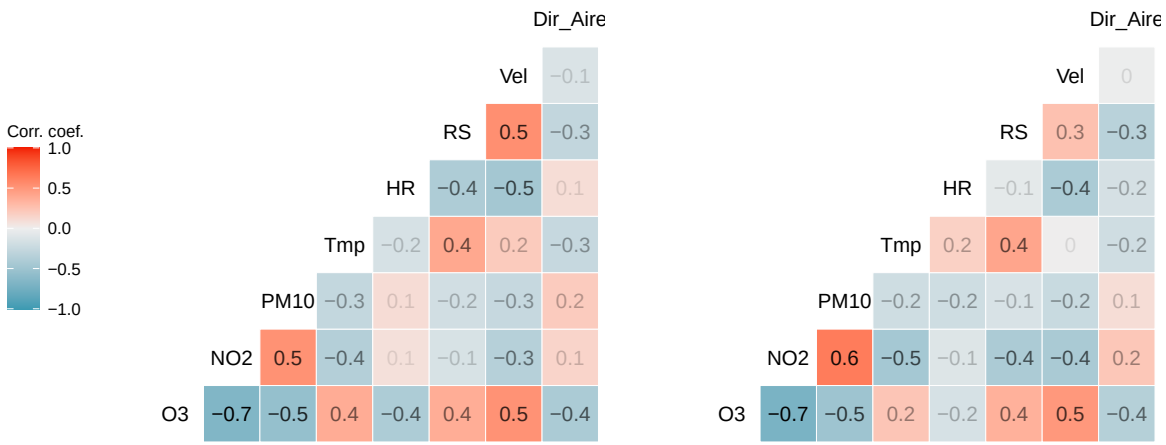
Si se utiliza la media diaria en vez de cada medición por separado, la matriz de correlaciones se muestra así:

```
air.corr.vars.daily <- air %>% dplyr::select(O3, NO2, PM10,
                                             Tmp, HR, RS, Vel, Dir_Aire,
                                             Fecha) %>%
  group_by(Fecha) %>% dplyr::summarise_all(mean, na.rm=TRUE) %>%
  dplyr::select(-Fecha)

cor(air.corr.vars.daily, method = 'pearson', use = 'complete.obs')
```

```
##              O3   NO2  PM10     Tmp     HR     RS     Vel Dir_Aire
## O3         1.00 -0.70 -0.49  0.2357 -0.150  0.352  0.4872   -0.355
## NO2       -0.70  1.00  0.64 -0.4503 -0.109 -0.409 -0.4209    0.226
## PM10      -0.49  0.64  1.00 -0.1980 -0.203 -0.137 -0.2370    0.101
## Tmp        0.24 -0.45 -0.20  1.0000  0.169  0.416  0.0031   -0.193
## HR        -0.15 -0.11 -0.20  0.1687  1.000 -0.063 -0.4148   -0.159
## RS         0.35 -0.41 -0.14  0.4161 -0.063  1.000  0.2612   -0.329
## Vel        0.49 -0.42 -0.24  0.0031 -0.415  0.261  1.0000   -0.015
## Dir_Aire -0.36  0.23  0.10 -0.1928 -0.159 -0.329 -0.0153    1.000
```

```
ggarrange(ncol=2, nrow=1, align='hv', common.legend=TRUE, legend='left',
  ggcorr(air.corr.vars, method=c('complete.obs', 'pearson'),
       name='Corr. coef.', label=TRUE, label_alpha=TRUE),
  ggcorr(air.corr.vars.daily, method=c('complete.obs', 'pearson'),
       name='Corr. coef.', label=TRUE, label_alpha=TRUE)) + title.centered +
  ggtitle('Matriz de correlaciones (mediciones - media diaria)')
```



Se observa que **las correlaciones bajan de intensidad en la mayoría de los casos**, aunque en algunos casos aumenta su intensidad. Se procede a analizar cada relación lineal que ve su intensidad modificada significativamente.

La concentración de O_3 en el ambiente influye menos en el aumento de temperaturas y en la bajada de la humedad relativa. También influye algo menos en la dirección del aire y la radiación solar.

La concentración de NO_2 influye más en el descenso de la radiación solar y la intensidad del viento; no ve su relación con el descenso de temperatura modificada.

Respecto a la concentración de partículas en suspensión, estas afectan algo menos al descenso de la temperatura, de la humedad relativa, de la radiación solar y de la velocidad del viento.

Es interesante analizar los cambios que sufre la correlación entre contaminantes y entre variables meteorológicas. Por ejemplo, la relación entre O_3 y NO_2, y la relación entre O_3 y PM_{10} se mantienen.

Se acentúa la correlación positiva entre la concentración de NO_2 y la de partículas en suspensión. Las correlaciones entre las variables meteorológicas han disminuido su intensidad, salvo en el caso de la relación entre temperatura y radiación solar.

1.2 Modelo de regresión lineal

1.2.1 Modelo de regresión lineal O_3~RS

Se desea explicar la cantidad de ozono en la atmósfera en función de la radiación solar mediante un modelo de regresión lineal. En el estudio de correlación lineal se ha observado que existe una correlación positiva moderada entre las 2 variables; por tanto se espera que el modelo no dé resultados muy buenos. Se procede a la creación del modelo:

```
rs.o3 <- lm(O3 ~ RS, data=air, na.action=na.omit)

summary(rs.o3)
```

```
##
## Call:
## lm(formula = O3 ~ RS, data = air, na.action = na.omit)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.59 -19.47   2.12  17.24  60.29
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 34.46439    0.34559    99.7 <0.0000000000000002 ***
## RS           0.05910    0.00173    34.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22 on 7250 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.138
## F-statistic: 1.16e+03 on 1 and 7250 DF,  p-value: <0.0000000000000002
```

$$\hat{O_3} = 34.36 + 0.05 \times RS$$

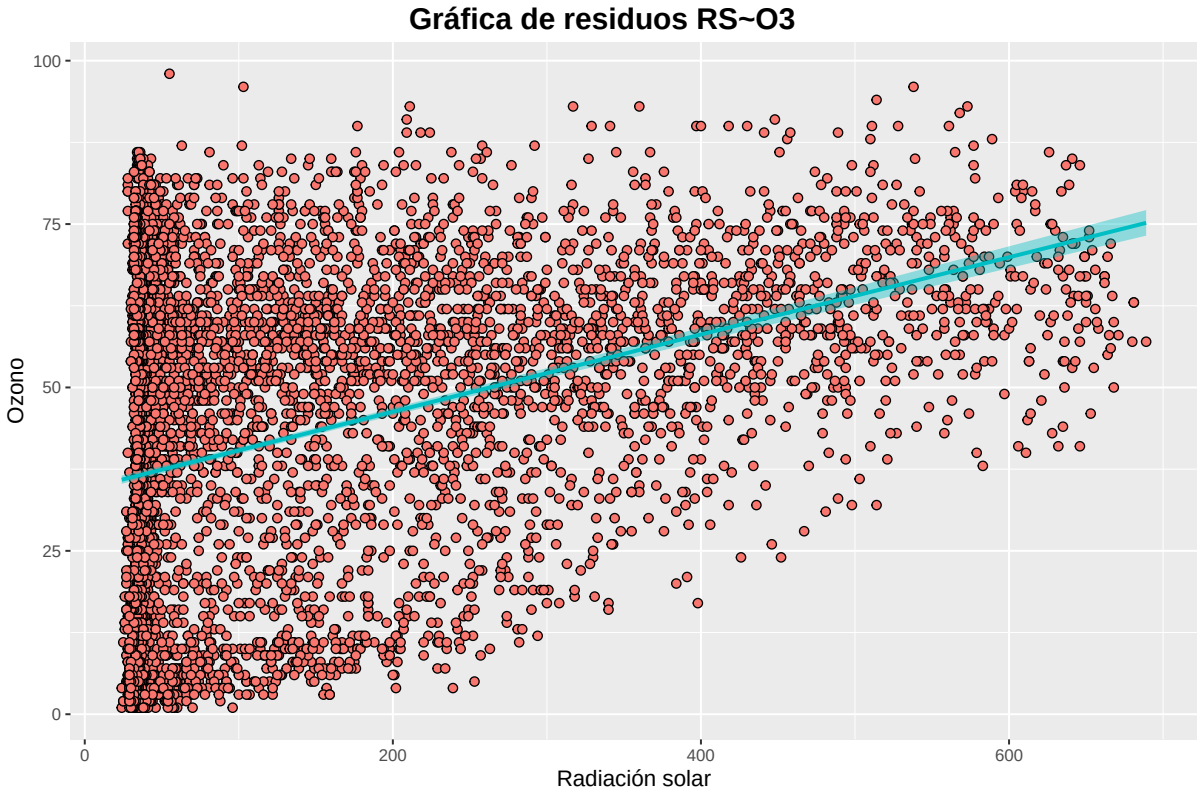
Tal y como se esperaba, el coeficiente de la radiación solar es positivo: esto indica que **a mayor radiación solar, mayor será la concentración de O_3 en la atmósfera**.

Para el coeficiente de radiación solar, el $p - value < 0.05$; por tanto con un nivel de significancia del 95% es estadísticamente significativo para el modelo.

Sin embargo, observando los valores de R^2 (coeficiente de determinación), el modelo solo explica un 14% de la variación de la concentración de Ozono en la atmósfera.

Teniendo un coeficiente de determinación bajo, se podría concluir que el modelo no es bueno. Sin embargo, también se tiene un coeficiente de radiación solar estadísticamente significativo, por lo que se podría obtener información interesante del modelo. Se procede a analizar la gráfica de residuos:

```
ggplot(data=air, mapping=aes(x=RS, y=O3)) +
  geom_point(size=2.0, shape=21, color='black', fill=default.color.main) +
  geom_smooth(formula=y~x, method='lm', color=default.color.secondary,
              fill=default.color.secondary) +
  title.centered + ggtitle('Gráfica de residuos RS~O3') +
  xlab('Radiación solar') + ylab('Ozono')
```



Se observa la correlación positiva entre la radiación solar y la cantidad de O_3 en el ambiente; sin embargo, la radiación solar por sí sola no puede explicar la variación de O_3. **El modelo no es adecuado**.

1.2.2 Modelo de regresión lineal O_3 ~ PM_{10}_cat

Se desea explicar la concentración de O_3 en función de la calidad del aire según el marcador de partículas en suspensión mediante un modelo de regresión lineal. En el estudio de correlación lineal se ha observado que existe una correlación negativa moderada entre las 2 variables; cuantas más partículas en suspensión existen en el aire, menos Ozono hay. No se esperan resultados muy buenos de este modelo.

En primer lugar, se procede a transformar la variable PM_{10} de acuerdo a la calidad del aire:

```
air$PM10_cat <- cut(air$PM10, breaks=c(0, 40, 60, 120, 160, 800),
                    labels=c('Muy buena', 'Buena', 'Mejorable',
                             'Mala', 'Muy mala'))
```

Patricia Lázaro Tello
plazarotello@uoc.edu

```
air %>% dplyr::select(PM10, PM10_cat) %>% head(n=5)
```

```
##   PM10  PM10_cat
## 1   95 Mejorable
## 2   80 Mejorable
## 3   38 Muy buena
## 4   36 Muy buena
## 5   30 Muy buena
```

```
summary(air$PM10_cat)
```

```
## Muy buena     Buena Mejorable      Mala  Muy mala
##      5254       670       820       237       271
```

Una vez categorizada la variable PM_{10}, se procede a crear el modelo de regresión lineal:

```
pm10cat.o3 <- lm(O3 ~ PM10_cat, data=air, na.action=na.omit)

summary(pm10cat.o3)
```

```
##
## Call:
## lm(formula = O3 ~ PM10_cat, data = air, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.59 -10.59   0.41  11.41  65.95
##
## Coefficients:
##                   Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         51.586      0.244   211.0 <0.0000000000000002 ***
## PM10_catBuena      -22.264      0.727   -30.6 <0.0000000000000002 ***
## PM10_catMejorable  -36.613      0.665   -55.0 <0.0000000000000002 ***
## PM10_catMala       -44.152      1.177   -37.5 <0.0000000000000002 ***
## PM10_catMuy mala   -44.535      1.104   -40.3 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18 on 7247 degrees of freedom
## Multiple R-squared:  0.442,  Adjusted R-squared:  0.442
## F-statistic: 1.44e+03 on 4 and 7247 DF,  p-value: <0.0000000000000002
```

$$\hat{O_3} = 51.586 + \begin{cases} 0 & \text{si la calidad de } PM_{10} \text{ es muy buena} \\ -22.264 & \text{si la calidad de } PM_{10} \text{ es buena} \\ -36.613 & \text{si la calidad de } PM_{10} \text{ es mejorable} \\ -44.152 & \text{si la calidad de } PM_{10} \text{ es mala} \\ -44.535 & \text{si la calidad de } PM_{10} \text{ es muy mala} \end{cases}$$

Al tratarse de una variables categórica, los coeficientes no pueden interpretarse de la misma manera que se interpretaban para variables continuas.

Se ha tomado como nivel de referencia una calidad del aire basada en partículas en suspensión muy buena. En este caso, **cuanto peor es la calidad del aire basada en esta medida, menos O_3 hay en el ambiente**.

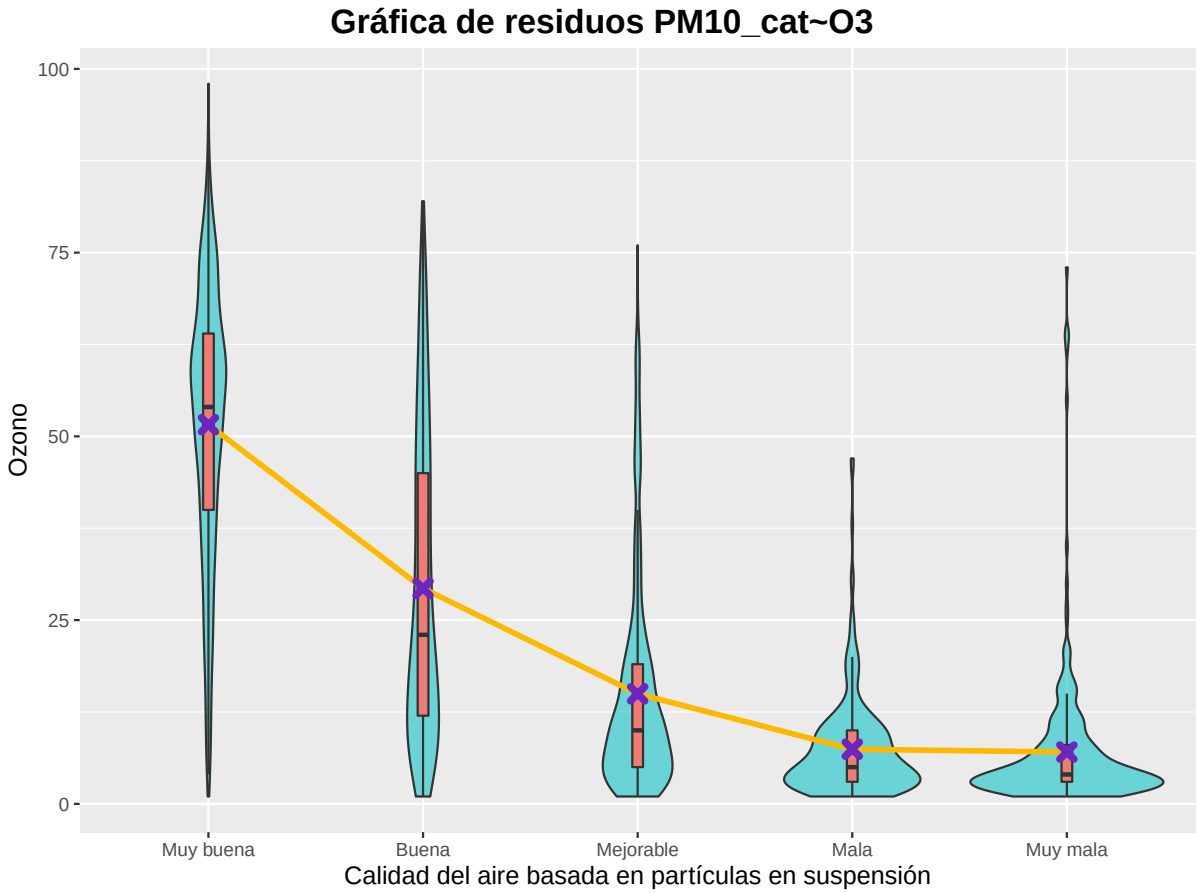
Traduciendo la variable categórica a su equivalente continuo, se puede afirmar que **cuantas más partículas en suspensión existan en el ambiente, menos O_3 habrá**. Esto concuerda con la correlación negativa observada inicialmente.

Además, el $p-value < 0.05$ para todos los coeficientes; con un nivel de significancia del 95%, el predictor es estadísticamente significativo para el modelo.

Al analizar los valores de R^2 (coeficiente de determinación), se obtiene que el modelo explica un 44% de la variación de la concentración de Ozono en la atmósfera. Esta variación es mucho mayor que la explicada en el modelo anterior, pero todavía es bastante baja. **El modelo no es adecuado**.

```
air.tmp <- air %>% dplyr::select(PM10_cat, O3) %>% group_by(PM10_cat) %>%
  summarize(avg=mean(O3))

ggplot(data=air, mapping=aes(x=PM10_cat, y=O3)) +
  geom_violin(fill=default.color.secondary, alpha=0.55) +
  geom_boxplot(width=0.05, outlier.shape=NA,
               fill=default.color.main, alpha=0.95) +
  xlab('Calidad del aire basada en partículas en suspensión') + ylab('Ozono') +
  ggtitle('Gráfica de residuos PM10_cat~O3') + title.centered +
  geom_line(data=air.tmp, mapping=aes(group=1, x=PM10_cat, y=avg), size=1.25,
            color=default.color.cinq) +
  geom_point(data=air.tmp, mapping=aes(x=PM10_cat, y=avg), shape=4, stroke=2.5,
             color=default.color.quat)
```



1.3 Modelo de regresión lineal múltiple

Se desea explicar la concentración de O_3 en función de la radiación solar, la concentración de NO_2, la temperatura y la dirección del aire mediante un modelo de regresión lineal múltiple.

En el estudio de correlación lineal se ha observado que existe una correlación positiva entre la concentración de O_3 y la radiación solar y la temperatura; a su vez, también existe una correlación negativa entre la cantidad de O_3 y la de NO_2 y la dirección del viento. Se espera que el modelo resultante sea adecuado.

Para seleccionar las variables que finalmente formarán parte del modelo final se va a seguir la estrategia de **selección hacia adelante** comprobando la bondad del ajuste tras cada expansión del modelo.

Se procede a tomar el modelo RS~O_3, expandirlo añadiendo el predictor Dir_Aire y

Patricia Lázaro Tello
plazarotello@uoc.edu

comprobar si el modelo ha mejorado. Por tanto, tendremos los siguientes modelos:

$$modelo_1 = \beta_0 + \beta_1 RS$$
$$modelo_2 = \beta_0 + \beta_1 RS + \beta_2 Dir\_Aire$$

```
modelo.1 <- rs.o3
summary(modelo.1)
```

```
##
## Call:
## lm(formula = O3 ~ RS, data = air, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.59 -19.47   2.12  17.24  60.29
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 34.46439    0.34559    99.7 <0.0000000000000002 ***
## RS           0.05910    0.00173    34.1 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22 on 7250 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.138
## F-statistic: 1.16e+03 on 1 and 7250 DF,  p-value: <0.0000000000000002
```

```
modelo.2 <- lm(O3 ~ RS+Dir_Aire, data=air, na.action=na.omit)
summary(modelo.2)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire, data = air, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.64 -17.02   0.93  15.60  57.74
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 50.98284    0.60983    83.6 <0.0000000000000002 ***
## RS           0.04420    0.00169    26.2 <0.0000000000000002 ***
## Dir_Aire    -0.07975    0.00250   -32.0 <0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21 on 7249 degrees of freedom
## Multiple R-squared:  0.245,  Adjusted R-squared:  0.244
## F-statistic: 1.17e+03 on 2 and 7249 DF,  p-value: <0.0000000000000002
```

En el modelo completo (modelo$_2$), se obtiene para todos los coeficientes $p-value < 0.05$; por tanto, todos los coeficientes son estadísticamente significativos con un nivel de significancia del 95%.

El predictor de la dirección del aire tiene un coeficiente de signo negativo, tal y como sugería la correlación negativa entre las variables O_3 y Dir_Aire.

El coeficiente de determinación del modelo R^2 del modelo completo es de 0,24, que frente al coeficiente de determinación del modelo reducido 0,14 indica que el modelo completo explica un 11% más de la variación de la concentración de Ozono en el ambiente.

El modelo ha mejorado, aunque todavía ha de mejorar más antes de poder decir que es un modelo adecuado para explicar la concentración de ozono en el ambiente.

Se procede a añadir al modelo anterior (modelo$_2$) la variable predictora NO_2, obteniéndose así el modelo$_3$.

$$modelo_3 = \beta_0 + \beta_1 RS + \beta_2 Dir\_Aire + \beta_3 NO_2$$

```
modelo.3 <- lm(O3 ~ RS+Dir_Aire+NO2, data=air, na.action=na.omit)
summary(modelo.3)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire + NO2, data = air, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -51.5  -11.3    0.0   10.8   56.5
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 67.89815    0.50530   134.4 <0.0000000000000002 ***
## RS           0.03395    0.00126    26.9 <0.0000000000000002 ***
## Dir_Aire    -0.06315    0.00187   -33.8 <0.0000000000000002 ***
## NO2         -1.43679    0.01883   -76.3 <0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 7248 degrees of freedom
## Multiple R-squared:  0.581,  Adjusted R-squared:  0.581
## F-statistic: 3.35e+03 on 3 and 7248 DF,  p-value: <0.0000000000000002
```

El nuevo modelo reducido es el modelo$_2$, mientras que el nuevo modelo completo es el modelo$_3$. Se procede a analizar el modelo completo y compararlo con el modelo reducido.

Observando los $p - value$ de los coeficientes, se obtiene que para todos ellos $p - value < 0.05$; por tanto, con un nivel de significancia del 95% todos los coeficientes son estadísticamente significativos para el modelo completo.

Como se observaba en el análisis de correlaciones lineales, las variables O_3 y NO_2 están fuertemente correladas de forma negativa. Esto se refleja en el coeficiente del predictor NO_2, de signo negativo y con un valor absoluto superior al resto de los predictores (salvo el valor absoluto de la constante).

El coeficiente de determinación del modelo R^2 del modelo completo es de 0,58, que frente al coeficiente de determinación del modelo reducido 0,24 indica que el modelo completo explica un 34% más de la variación de la concentración de Ozono en el ambiente, hasta llegar a explicar un 58% de la variación.

El modelo completo mejora significativamente al añadir la concentración de NO_2 como predictor; ya está muy cerca de considerarse un modelo adecuado.

Se procede a añadir la variable temperatura (Tmp) al modelo, generando el nuevo modelo completo modelo$_4$ (el nuevo modelo reducido es modelo$_3$).

$$modelo_4 = \beta_0 + \beta_1 RS + \beta_2 Dir\_Aire + \beta_3 NO_2 + \beta_4 Tmp$$

Existe una correlación positiva moderada entre la concentración de O_3 y la temperatura ($corr(O_3, Tmp) = 0.4$). No se espera que la variable aporte mucho al modelo ya existente.

Además, se observa que existe correlación moderada positiva entre la radiación solar y la variación de temperatura. Se procede a explorar la colinealidad de estas 2 variables para decidir si introducir la temperatura como predictor.

```
faraway::vif(dplyr::select(air, RS, Tmp))
```

```
##  RS Tmp
## 1.2 1.2
```

Un factor de inflación de la varianza (FIV) superior a 5 implica que existe una colinealidad fuerte entre ambas variables; es decir, que no son independientes. En este caso, se asume que no existe colinealidad entre las variables. Se procede a incluir la temperatura en el modelo:

```
modelo.4 <- lm(O3 ~ RS+Dir_Aire+NO2+Tmp, data=air, na.action=na.omit)
summary(modelo.4)
```

```
##
## Call:
## lm(formula = O3 ~ RS + Dir_Aire + NO2 + Tmp, data = air, na.action = na.omit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51.52 -11.34   0.08  10.79  56.08
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 66.63203    0.91250   73.02 <0.0000000000000002 ***
## RS           0.03318    0.00135   24.63 <0.0000000000000002 ***
## Dir_Aire    -0.06274    0.00189  -33.26 <0.0000000000000002 ***
## NO2         -1.42606    0.01989  -71.68 <0.0000000000000002 ***
## Tmp          0.07880    0.04729    1.67                0.096 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 7247 degrees of freedom
## Multiple R-squared:  0.581,  Adjusted R-squared:  0.581
## F-statistic: 2.52e+03 on 4 and 7247 DF,  p-value: <0.0000000000000002
```

Observando los $p - value$ de los coeficientes, se obtiene que para el coeficiente de temperatura $p - value = 0.096 > 0.05$; por tanto, con un nivel de significancia del 95% el coeficiente de temperatura no es estadísticamente significativo para el modelo completo.

El coeficiente del predictor temperatura tiene signo positivo y un valor pequeño; esto concuerda con la correlación débil y positiva comentada anteriormente.

El coeficiente de determinación del modelo R^2 del modelo completo es de 0,58, que

frente al coeficiente de determinación del modelo reducido 0,58 indica que el modelo completo explica un 0,01% más de la variación de la concentración de Ozono en el ambiente.

En conclusión, pese a que el Factor de Inflación de la Varianza recomendaba introducir la variable temperatura en el modelo por no mantener colinealidad con la radiación solar, su inclusión no ha producido ninguna mejora sobre el modelo reducido $modelo_3$ y el predictor Tmp no es significativo.

Estos resultados ofrecen la noción de que **temperatura y radiación solar se encuentran relacionados, pero no de manera lineal**.

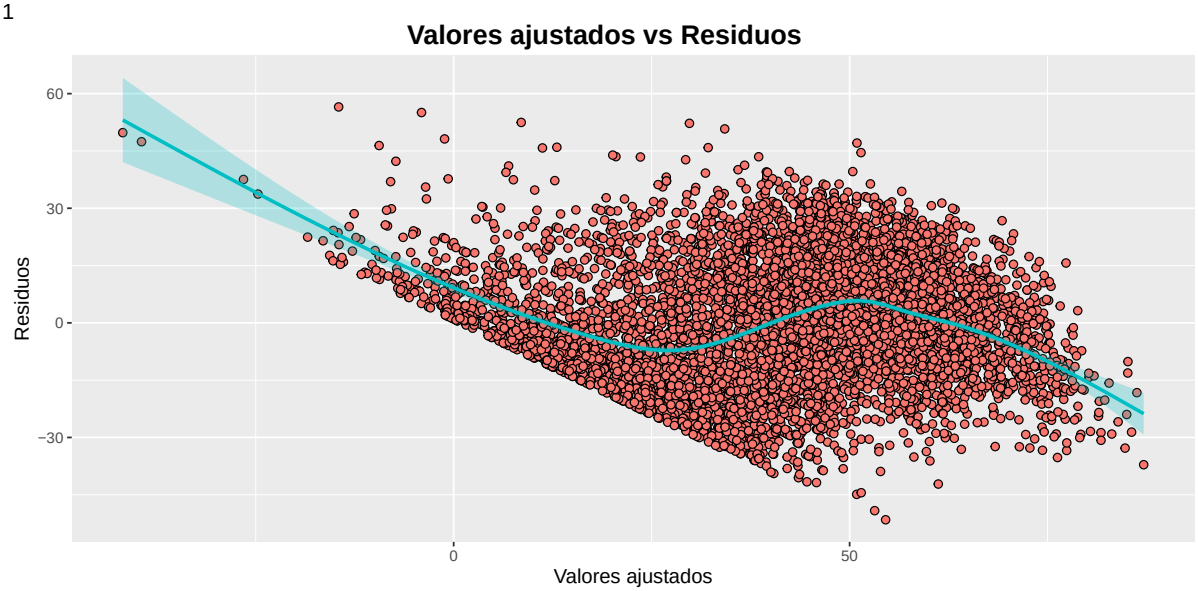
1.4 Diagnosis del modelo

En el diagnóstico del modelo ($modelo_3$) se va a comprobar si los residuos tienen varianza constante (**homocedasticidad**) y se distribuyen de forma normal. Se procede a construir las gráficas para el diagnóstico:

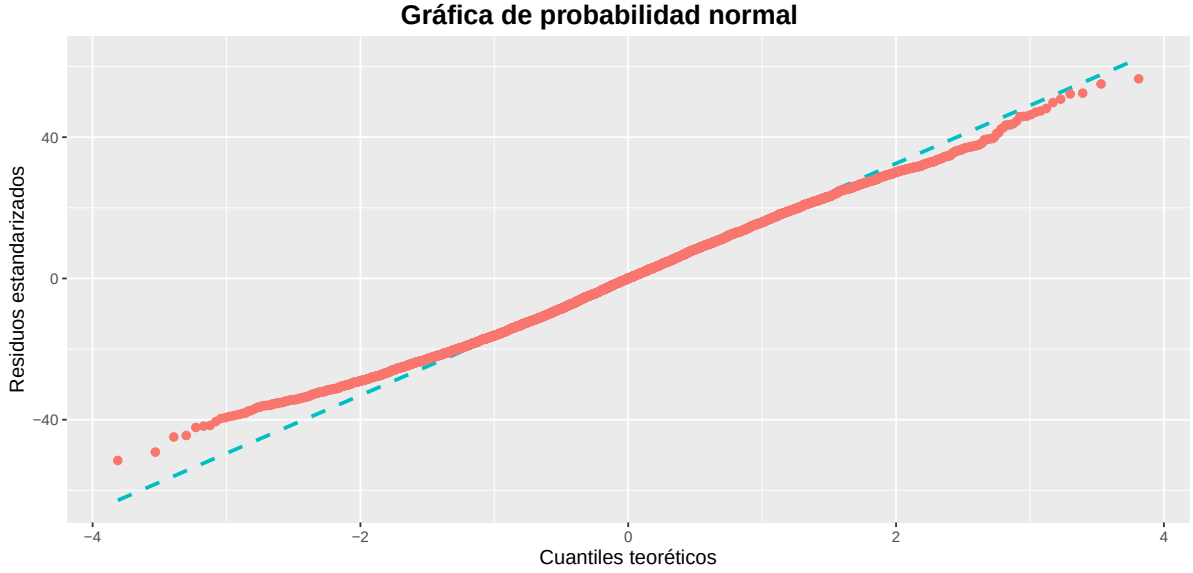
```
modelo.3.fortified <- fortify(modelo.3)
ggarrange(nrow=2, ncol=1,
  ggplot(modelo.3.fortified, aes(x=.fitted, y=.resid)) +
    geom_point(size=2, shape=21, fill=default.color.main) +
    geom_smooth(color=default.color.secondary,
                fill=default.color.secondary, alpha=0.25) +
    title.centered + ggtitle('Valores ajustados vs Residuos') +
    xlab('Valores ajustados') + ylab('Residuos') + labs(tag='1'),

  ggplot(modelo.3.fortified, aes(sample=.resid)) +
    geom_qq_line(linetype='dashed', size=1.05, color=default.color.secondary) +
    geom_qq(size=2, color=default.color.main) + title.centered +
    ggtitle('Gráfica de probabilidad normal') + labs(tag='2') +
    xlab('Cuantiles teoréticos') + ylab('Residuos estandarizados'))
```

1



2



El primer gráfico se utiliza para comprobar que la varianza del error sea constante; analizándolo se observa que no es así. Los datos presentan un patrón reconocible: los datos están acotados por una línea oblícua descendente en el lado inferior izquierdo. Se puede afirmar que **la varianza del error no es constante**.

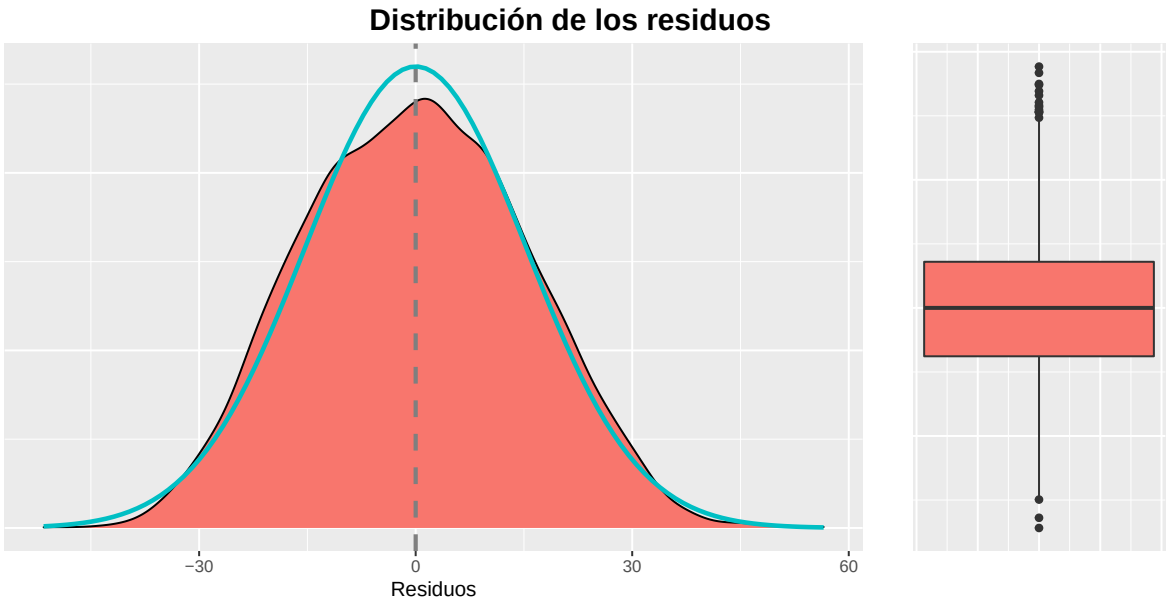
Respecto a la gráfica QQ o gráfica de probabilidad normal de los errores, se observan colas algo más pesadas que las de la distribución normal. Se procede a observar la distribución de los valores respecto a la distribución normal y aplicar un test de normalidad:

```
lillie.test(modelo.3.fortified$.resid)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  modelo.3.fortified$.resid
## D = 0.02, p-value = 0.00000007
```

```
ggarrange(ncol=2, nrow=1, widths=c(3,1), align='hv',
  ggplot(modelo.3.fortified, aes(x=.resid)) +
    geom_density(mapping=aes(y=..density..), fill=default.color.main) +
    geom_vline(xintercept=mean(modelo.3.fortified$.stdresid), size=1.05,
               linetype='dashed', color='gray50') +
    stat_function(fun=dnorm, args=c(mean=mean(modelo.3.fortified$.resid),
                                    sd=sd(modelo.3.fortified$.resid)),
                  color=default.color.secondary, size=1.15) +
    no.axis.y + xlab('Residuos') + ylab(''),

  ggplot(modelo.3.fortified, aes(x=.resid)) + coord_flip() +
    geom_boxplot(fill=default.color.main) + no.axis.x + no.axis.y +
    xlab('') + ylab('')) +
  title.centered + ggtitle('Distribución de los residuos')
```



Se observa que los residuos podrían seguir a *grosso modo* una distribución normal de media 0 y desviación estándar 15. Los residuos se concentran más en las colas que

en el centro, provocando la colas pesadas que se observaban en el gráfico QQ.

Sin embargo, el test de Kolmogorov-Smirnov ofrece un $p - value < 0.05$, lo que quiere decir que, con un nivel de significancia del 95%, **los residuos no siguen una distribución normal**.

Por todo lo anterior, se puede concluir que el modelo no termina de ajustar bien a los datos y **la suposición de homocedasticidad y normalidad de residuos no se cumplen**.

1.5 Predicción del modelo

$$modelo_4 = \beta_0 + \beta_1 RS + \beta_2 Dir\_Aire + \beta_3 NO_2 + \beta_4 Tmp$$

Tomando el modelo completo (modelo$_4$), se busca predecir la concentración de O_3:

```
predict.vars <- data.frame(RS=180, NO2=15, Dir_Aire=250, Tmp=20)
predicted.o3 <- predict.lm(modelo.4, predict.vars)[[1]]
```

Se obtiene una concentración de O_3 predicha de 37.

2 Regresión logística

Se procede a crear el índice de calidad del aire basado en O_3 como factor (calidades buena y mejorable) y como variable numérica (calidad buena como 0 y calidad mejorable como 1):

```
air$icO3.factor <- cut(air$O3, breaks=c(0, 80, 100),
                       labels=c('buena', 'mejorable'))
air$icO3.factor[is.na(air$O3)] <- 'mejorable'
air$icO3 <- as.numeric(air$icO3.factor) - 1

air %>% dplyr::select(O3, icO3, icO3.factor) %>% head(n=5)
```

```
##   O3 icO3 icO3.factor
## 1  5    0       buena
## 2 10    0       buena
## 3 14    0       buena
## 4 11    0       buena
## 5 16    0       buena
```

2.1 Estudio de relaciones entre variables. Análisis crudo de posibles factores de riesgo

Se procede a categorizar las variables RS y Vel:

```
air$RS_cat2 <- cut(air$RS, breaks=c(0, 100, 700),
                   labels=c('normal_baja', 'normal_alta'))

air$Vel_cat2 <- cut(air$Vel, breaks=c(0, 3, 10),
                   labels=c('flojo', 'moderado'))

air %>% dplyr::select(RS, RS_cat2, Vel, Vel_cat2) %>% head(n=5)
```

```
##   RS     RS_cat2  Vel Vel_cat2
## 1 36 normal_baja 1.49    flojo
## 2 36 normal_baja 0.98    flojo
## 3 37 normal_baja 0.98    flojo
## 4 37 normal_baja 0.99    flojo
## 5 37 normal_baja 0.99    flojo
```

Se desea comprobar si existe asociación entre el índice de calidad del aire basado

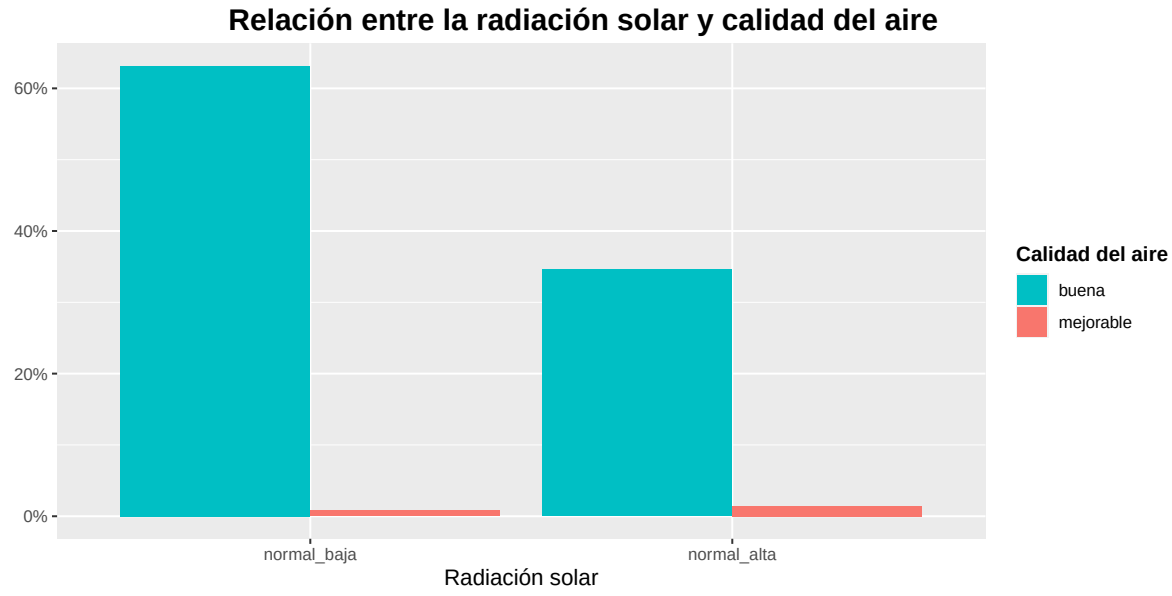
en O_3 y la radiación solar (categorizada), la velocidad del viento (categorizada) y la humedad del ambiente. Para ello se aplicará el test χ^2 de Pearson y se calcularán las OR (*Odds-Ratio*):

```
odds.ratio <- function(contigency){
  tmp.1 <- (contigency[1] / contigency[2])
  tmp.2 <- (contigency[3] / contigency[4])
  return(tmp.1/tmp.2)
}
```

A continuación se procede a evaluar el test χ^2 de Pearson y las *Odds-Ratio* de la calidad del aire basado en O_3 y la radiación solar, la velocidad del viento y la humedad del ambiente.

Respecto a la posible asociación de la calidad del aire basado en O_3 y la radiación solar:

```
ggplot(data=air, mapping=aes(x=RS_cat2, fill=ic03.factor,
                             ..count../sum(..count..))) +
  geom_bar(position='dodge') + guides(fill=guide_legend('Calidad del aire')) +
  xlab('Radiación solar') + ylab('') + title.centered +
  ggtitle('Relación entre la radiación solar y calidad del aire') +
  theme(legend.title=element_text(face='bold')) +
  scale_y_continuous(labels=scales::percent) +
  scale_fill_manual(values=palette[2:1])
```



```
chisq.test(table(air$ic03, air$RS_cat2))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(air$ic03, air$RS_cat2)
## X-squared = 51, df = 1, p-value = 0.0000000000008
```

```
odds.ratio(table(air$ic03, air$RS_cat2))
```

```
## [1] 3.1
```

Se observa que **existe mayor proporción de niveles de calidad del aire buenos cuando la radiación solar es normal/baja**; mientras que cuando la radiación solar es normal/alta, aumentan ligeramente las observaciones de calidad del aire mejorable.

Respecto al test de χ^2 de Pearson, se observa que $p - value = 0 < \alpha = 0.05$; por tanto, existe asociación entre las 2 variables.

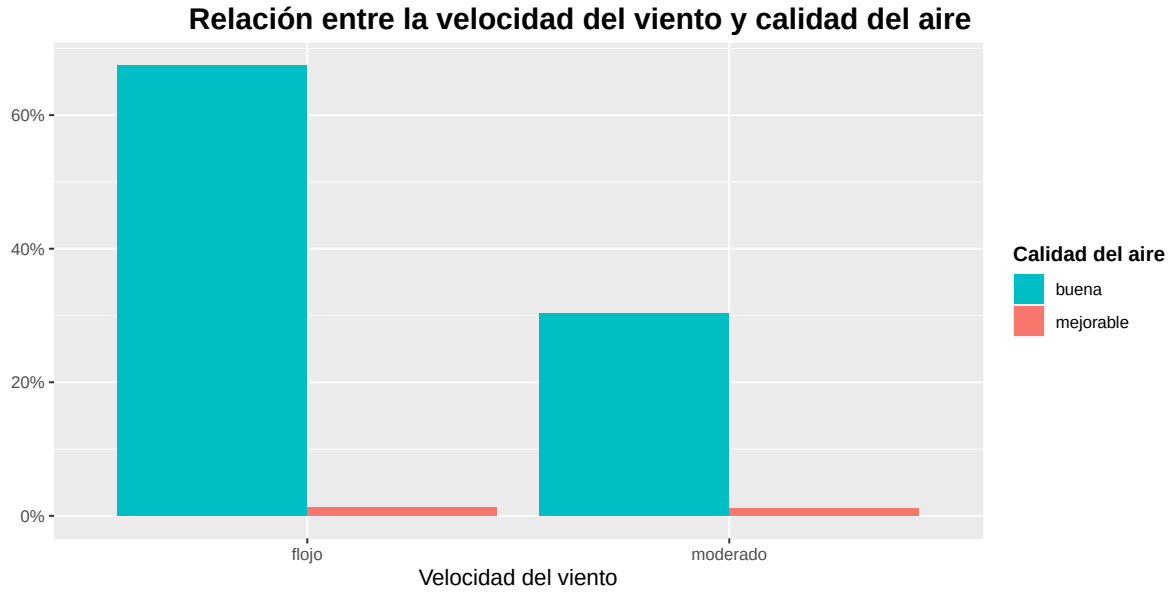
Por último, el valor de *Odds-Ratio* confirma que la asociación es tal que la radiación solar es un **factor de riesgo** para la calidad del aire basado en O_3.

Esto quiere decir que la probabilidad de encontrar una calidad del aire mejorable (no buena) cuando la radiación solar es normal/alta es de 3,1 veces respecto a encontrar una calidad del aire mejorable cuando la radiación solar es normal/baja.

Como conclusión, se observa que **la radiación solar es un factor de riesgo para la calidad del aire basado en O_3**.

Respecto a la posible asociación de la calidad del aire basado en O_3 y la velocidad del viento:

```
ggplot(data=air, mapping=aes(x=Vel_cat2, fill=ic03.factor,
                             ..count../sum(..count..))) +
  geom_bar(position='dodge') + guides(fill=guide_legend('Calidad del aire')) +
  xlab('Velocidad del viento') + ylab('') + title.centered +
  ggtitle('Relación entre la velocidad del viento y calidad del aire') +
  theme(legend.title=element_text(face='bold')) +
  scale_y_continuous(labels=scales::percent) +
  scale_fill_manual(values=palette[2:1])
```




```
chisq.test(table(air$ic03, air$Vel_cat2))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(air$ic03, air$Vel_cat2)
## X-squared = 17, df = 1, p-value = 0.00004
```

```
odds.ratio(table(air$ic03, air$Vel_cat2))
```

```
## [1] 1.9
```

En el gráfico se observa que existe mayor proporción de niveles de calidad del aire buenos cuando la velocidad del viento es floja; mientras que cuando la velocidad del viento es moderada, aumentan ligeramente las observaciones de calidad del aire mejorable.

Respecto al test de χ^2 de Pearson, se observa que $p - value = 0 < \alpha = 0.05$; por tanto, existe asociación entre las 2 variables.

Por último, el valor de *Odds-Ratio* confirma que la asociación es tal que la velocidad del viento es un factor de riesgo para la calidad del aire basado en O_3.

Esto quiere decir que la probabilidad de encontrar una calidad del aire mejorable (no buena) cuando la velocidad del viento es moderada es de 1,9 veces respecto a encontrar una calidad del aire mejorable cuando la velocidad del viento es baja (flojo).

Como conclusión, se observa que **la velocidad del viento es un factor de riesgo**

para la calidad del aire basado en O_3.

Respecto a la posible asociación de la calidad del aire basado en O_3 y la humedad relativa:

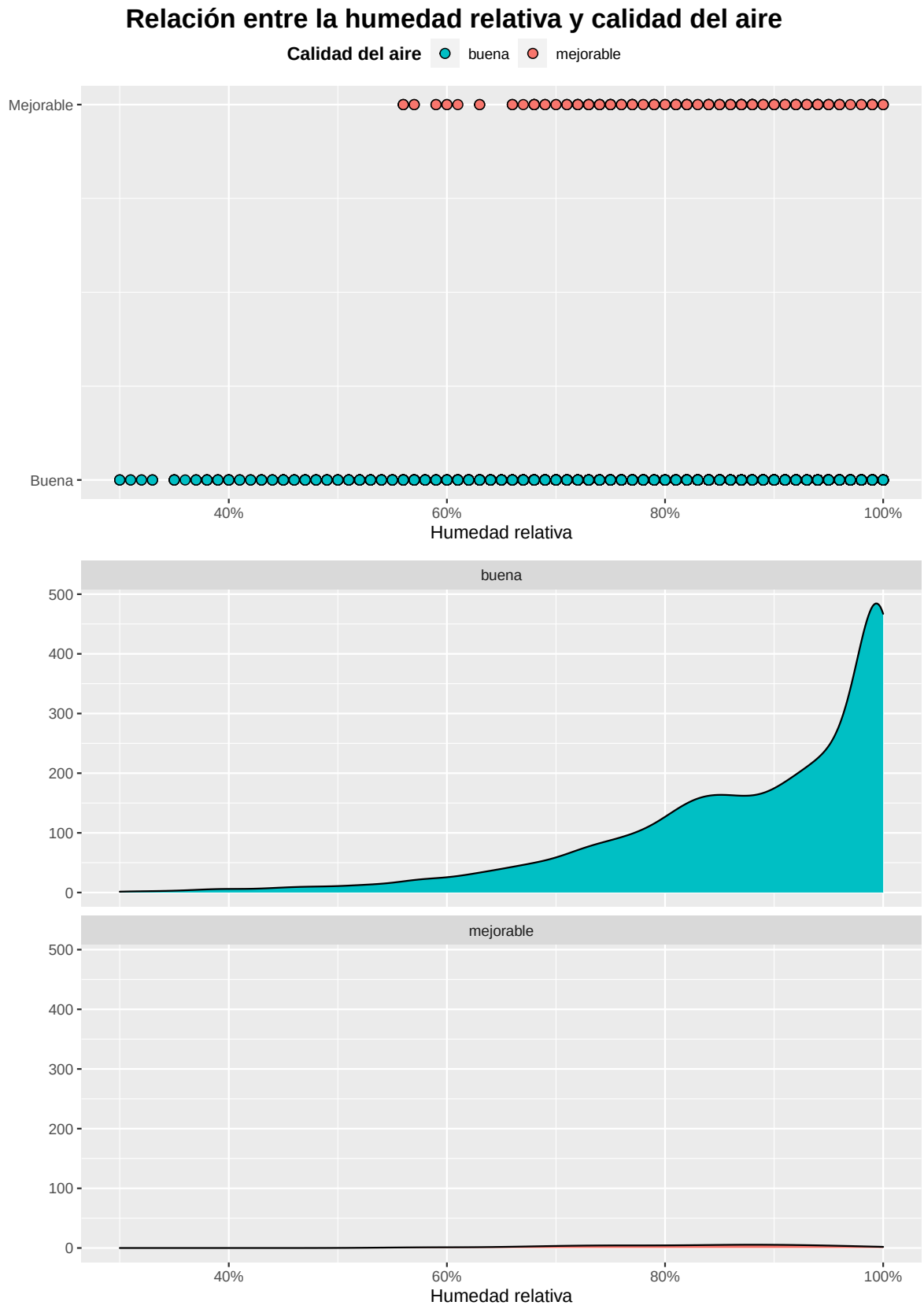
```
chisq.test(table(air$icO3, air$HR))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(air$icO3, air$HR)
## X-squared = 153, df = 69, p-value = 0.00000003
```

```
ggarrange(nrow=2, ncol=1, align='hv', common.legend=TRUE,
          heights=c(1.15,1.85),
  ggplot(data=air, mapping=aes(x=HR, fill=icO3.factor, y=icO3)) +
    geom_point(shape=21, size=2.5) +
    guides(fill=guide_legend('Calidad del aire')) +
    xlab('Humedad relativa') + ylab('') +
    theme(legend.title=element_text(face='bold')) +
    scale_x_continuous(labels=scales::percent_format(scale=1)) +
    scale_y_continuous(breaks=c(0.0, 1.0), minor_breaks=c(0.25, 0.5, 0.75),
                       labels=as_labeller(c(`0`='Buena', `1`='Mejorable'))) +
    scale_fill_manual(values=palette[2:1]),

  ggplot(data=air, mapping=aes(x=HR, fill=icO3.factor, y=..count..)) +
    geom_density() + facet_wrap(icO3.factor~., ncol=1, nrow=2) +
    guides(fill=guide_legend('Calidad del aire')) +
    xlab('Humedad relativa') + ylab('') +
    theme(legend.title=element_text(face='bold')) +
    scale_x_continuous(labels=scales::percent_format(scale=1)) +
    scale_fill_manual(values=palette[2:1])) +
  title.centered +
  ggtitle('Relación entre la humedad relativa y calidad del aire')
```

Patricia Lázaro Tello
plazarotello@uoc.edu



En el gráfico se observa que a menor humedad relativa en el ambiente, más posibilidades existen de que la calidad del aire sea buena.

Respecto al test de χ^2 de Pearson, se observa que $p - value = 0 < \alpha = 0.05$; por tanto, existe asociación entre las 2 variables.

La *Odds-Ratio* no es calculable, ya que la humedad relativa es continua. Sin embargo, observando el gráfica se puede intuir que la humedad relativa en el ambiente es un **factor de riesgo**: es más probable que la calidad del aire sea mejorable (no buena) cuando hay mucha humedad en el ambiente.

2.2 Modelo de regresión logística

Se procede a crear el modelo de regresión logística con la calidad del aire basado en O_3 como variable dependiente y la radiación solar categorizada como predictora.

$$modelo_1 : logit_1 = \beta_0 + \beta_1 RS$$

```
modelo.1 <- glm(icO3~RS_cat2, data=air, family=binomial)
summary(modelo.1)
```

```
##
## Call:
## glm(formula = icO3 ~ RS_cat2, family = binomial, data = air)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.284  -0.284  -0.163  -0.163   2.943
##
## Coefficients:
##                    Estimate Std. Error z value             Pr(>|z|)
## (Intercept)          -4.319      0.129  -33.51 < 0.0000000000000002 ***
## RS_cat2normal_alta    1.127      0.163    6.89      0.0000000000055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1567.1  on 7251  degrees of freedom
## Residual deviance: 1517.5  on 7250  degrees of freedom
## AIC: 1521
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
exp(confint(modelo.1))
```

```
##                    2.5 % 97.5 %
## (Intercept)         0.01  0.017
## RS_cat2normal_alta  2.25  4.271
```

Del resumen del modelo se pueden extraer varias conclusiones. En primer lugar, se observa que el predictor de radiación solar es estadísticamente significativo ($p-value < 0.05$) con un nivel de significancia del 95%.

A su vez, su coeficiente para la radiación normal/alta es positivo, lo que indica que las posibilidades de que la calidad del aire sea mejorable es mayor si la radiación es normal/alta que si es normal/baja. De esto se concluye que al aumentar la radiación solar hay más posibilidades de encontrar una calidad del aire basado en O_3 mejorable (no buena).

La *devianza* nula es superior a la *devianza* residual; por lo tanto, la radiación solar mejora el modelo.

Por último, se estima que el *Odds-Ratio* es (2.25, 4.271) con una significancia del 95%. Esto significa que la probabilidad de encontrar una calidad del aire mejorable es entre 2.25 y 4.271 veces mayor si la radiación solar es normal/alta que si es normal/baja.

Como el OR es mayor que 1, se confirma que **la radiación solar es un factor de riesgo para la calidad del aire basado en O_3**.

A continuación, se plantea añadir al $modelo_1$ la variable predictora temperatura.

$$modelo_2 : logit_2 = \beta_0 + \beta_1 RS + \beta_2 Tmp$$

```
modelo.2 <- glm(ico3~RS_cat2+Tmp, data=air, family=binomial)
summary(modelo.2)
```

```
##
## Call:
## glm(formula = ico3 ~ RS_cat2 + Tmp, family = binomial, data = air)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.448  -0.244  -0.185  -0.143   3.139
```

```
##
## Coefficients:
##                    Estimate Std. Error z value             Pr(>|z|)
## (Intercept)         -5.8428     0.3632  -16.09 < 0.0000000000000002 ***
## RS_cat2normal_alta   0.7895     0.1763    4.48            0.0000076 ***
## Tmp                  0.1063     0.0224    4.75            0.0000020 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1567.1  on 7251  degrees of freedom
## Residual deviance: 1492.8  on 7249  degrees of freedom
## AIC: 1499
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la nueva predictora, la variable temperatura, es estadísticamente significativa ($p-value < 0.05$) con un nivel de significancia del 95%, y que su coeficiente es positivo, por lo que **un aumento de la temperatura supone un aumento de las posibilidades de encontrar una calidad del aire mejorable**.

A su vez, el criterio de información de Akaike (AIC) del modelo$_2$ es inferior al del modelo$_1$; por tanto, ajusta mejor al modelo. Su *devianza* residual también es menor que su *devianza* nula, por lo que se concluye que **el nuevo predictor mejora el modelo**.

La adición de la variable predictora temperatura al modelo supone un cambio en el coeficiente de la radiación solar, que pasa de 1.127 a 0.7895, una bajada del 30%. **La variable temperatura es una variable de confusión**.

A continuación se decide añadir la variable humedad relativa al modelo$_1$, obteniéndose el modelo$_3$:

$$modelo_3 : logit_3 = \beta_0 + \beta_1 RS + \beta_2 HR$$

```
modelo.3 <- glm(ico3~RS_cat2+HR, data=air, family=binomial)
summary(modelo.3)
```

```
##
## Call:
## glm(formula = ico3 ~ RS_cat2 + HR, family = binomial, data = air)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.468  -0.257  -0.167  -0.149   3.011
##
## Coefficients:
##                    Estimate Std. Error z value   Pr(>|z|)
## (Intercept)        -2.44170    0.51141   -4.77 0.00000180 ***
## RS_cat2normal_alta  0.90729    0.17396    5.22 0.00000018 ***
## HR                 -0.02079    0.00559   -3.72     0.0002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1567.1  on 7251  degrees of freedom
## Residual deviance: 1504.8  on 7249  degrees of freedom
## AIC: 1511
##
## Number of Fisher Scoring iterations: 7
```

Se observa que la nueva predictora, la variable humedad, es estadísticamente significativa ($p - value < 0.05$) con un nivel de significancia del 95%, y que su coeficiente es negativo, por lo que un descenso de la humedad supone un aumento de las posibilidades de encontrar una calidad del aire mejorable.

A su vez, el criterio de información de Akaike (AIC) del modelo_3 es inferior al del modelo_1; por tanto, ajusta mejor al modelo. Su *devianza* residual también es menor que su *devianza* nula, por lo que se concluye que **el nuevo predictor mejora el modelo**.

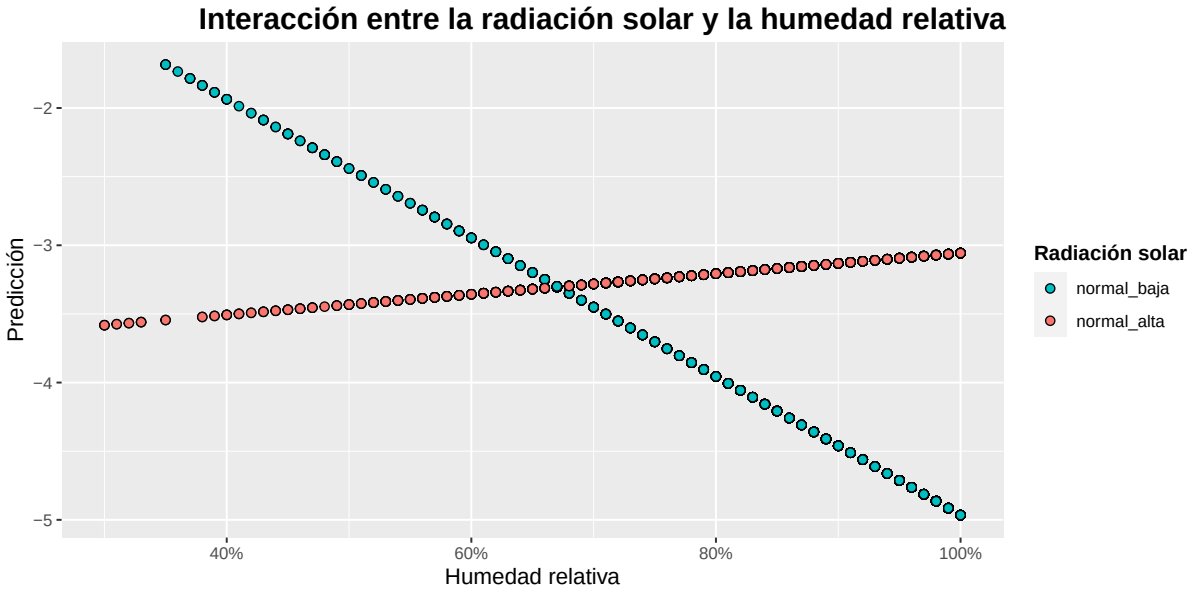
```
modelo.3.1 <- glm(ico3~RS_cat2+HR+RS_cat2:HR, data=air, family=binomial)
summary(modelo.3.1)
```

```
##
## Call:
## glm(formula = ico3 ~ RS_cat2 + HR + RS_cat2:HR, family = binomial,
##     data = air)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.584  -0.279  -0.159  -0.121   3.154
##
## Coefficients:
##                       Estimate Std. Error z value         Pr(>|z|)
## (Intercept)            0.08297    0.57420    0.14             0.89
```

```
## RS_cat2normal_alta    -3.89019    0.88540   -4.39 0.00001114264234 ***
## HR                    -0.05048    0.00691   -7.31 0.00000000000027 ***
## RS_cat2normal_alta:HR  0.05799    0.01063    5.46 0.00000004837756 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1567.1  on 7251  degrees of freedom
## Residual deviance: 1475.4  on 7248  degrees of freedom
## AIC: 1483
##
## Number of Fisher Scoring iterations: 7
```

```
modelo.3.1.predictions <- predict(modelo.3.1, type='link')

ggplot(data=air, mapping=aes(x=HR, fill=RS_cat2, y=modelo.3.1.predictions)) +
  geom_point(shape=21, size=2) + title.centered +
  guides(fill=guide_legend('Radiación solar')) +
  xlab('Humedad relativa') + ylab('Predicción') +
  theme(legend.title=element_text(face='bold')) +
  scale_x_continuous(labels=scales::percent_format(scale=1)) +
  scale_fill_manual(values=palette[2:1]) +
  ggtitle('Interacción entre la radiación solar y la humedad relativa')
```



Existe interacción entre los predictores humedad relativa y radiación solar

($p - value < 0.05$) con nivel de significancia del 95%. Además, se observa en la gráfica que las rectas no son paralelas; es decir, que la humedad relativa modifica el efecto de la radiación solar.

Por último, se decide crear un modelo$_4$ con las variables explicativas radiación solar y dirección del viento:

$$modelo_4 : logit_4 = \beta_0 + \beta_1 RS + \beta_2 Dir\_Aire$$

```
modelo.4 <- glm(icO3~RS_cat2+Dir_Aire, data=air, family=binomial)
summary(modelo.4)
```

```
##
## Call:
## glm(formula = icO3 ~ RS_cat2 + Dir_Aire, family = binomial, data = air)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.481  -0.256  -0.129  -0.095   3.554
##
## Coefficients:
##                    Estimate Std. Error z value             Pr(>|z|)
## (Intercept)        -2.76945    0.17830  -15.53 < 0.0000000000000002 ***
## RS_cat2normal_alta  0.68038    0.16772    4.06              0.00005 ***
## Dir_Aire           -0.01002    0.00111   -9.06 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1567.1  on 7251  degrees of freedom
## Residual deviance: 1404.9  on 7249  degrees of freedom
## AIC: 1411
##
## Number of Fisher Scoring iterations: 8
```

Se observa que la nueva predictora, la variable dirección del aire, es estadísticamente significativa ($p - value < 0.05$) con un nivel de significancia del 95%, y que su coeficiente es negativo, por lo que **un descenso de la dirección del aire (es decir, que el aire sople del noreste) supone un aumento de las posibilidades de encontrar una calidad del aire mejorable**.

A su vez, el criterio de información de Akaike (AIC) del modelo$_4$ es inferior al del

modelo$_1$; ajusta mejor al modelo. Su *devianza* residual también es menor que su *devianza* nula, por lo que se concluye que **el nuevo predictor mejora el modelo**.

2.3 Predicción

Se desea predecir la probabilidad de que la concentración de O_3 sea o no superior a 80 (es decir, que la calidad del aire basado en O_3 sea mejorable) dado los siguientes datos y el modelo$_4$:

```
predict.vars <- data.frame(RS_cat2='normal_alta', Dir_Aire=40)
predicted.o3 <- predict(modelo.4, predict.vars, type='response')[[1]]
```

La probabilidad predicha de que la calidad del aire basado en O_3 sea mejorable es de un 7,7%; si se ajusta el umbral de clasificación al 50%, se obtiene que en este caso **la calidad del aire predicha es buena**.

2.4 Bondad del ajuste

A continuación se procede a calcular la bondad del ajuste del modelo$_4$ mediante el test de Hosmer-Lemeshow:

```
hoslem.test(air$icO3, fitted(modelo.4))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  air$icO3, fitted(modelo.4)
## X-squared = 32, df = 8, p-value = 0.00008
```

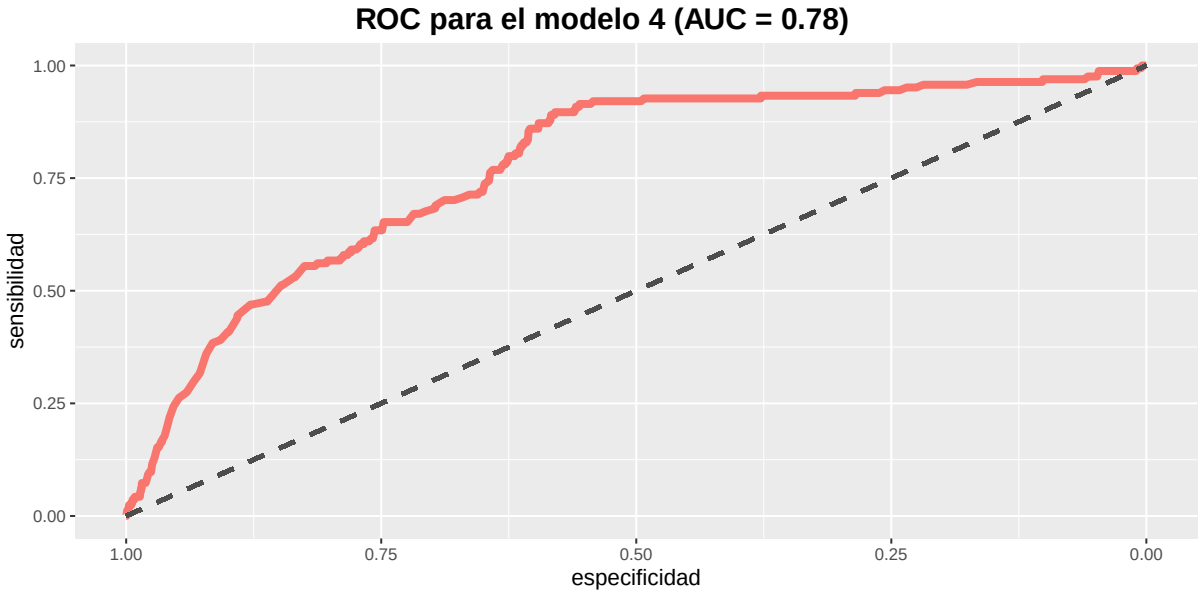
La hipótesis nula de este test supone que no hay diferencias entre los valores esperados y los valores observados. Como $p - value < 0.05$, con un nivel de significancia del 95% se puede afirmar que **el modelo no ajusta bien a los datos**.

2.5 Curva ROC

A continuación se procede a realizar un análisis ROC del modelo$_4$:

```
prob <- predict(modelo.4, type='response')
r <- roc(air$ic03, prob, data=air)

ggroc(r, color=default.color.main, size=2) +
  geom_segment(aes(x=1, xend=0, y=0, yend=1), color='grey30',
               linetype='dashed', size=1.05) +
  title.centered + ggtitle(paste0('ROC para el modelo 4 ',
                                  '(AUC = ', round(auc(r),2), ')')) +
  xlab('especificidad') + ylab('sensibilidad')
```



Se observa que el área bajo la curva es pequeña ($0.6 < 0,78 < 0.8$); por tanto, el modelo no discrimina muy bien los datos, tal y como ya sugería el test de Hosmer-Lemeshow.

Además, analizando el gráfico de la curva se observa que el umbral de clasificación más adecuado para el modelo se sitúa muy cerca de (0.5, 1.0). Esto sugiere que la probabilidad de de que la calidad del aire sea mejorable no puede ser discriminada de forma lineal; es decir, unas probabilidades muy altas de que la calidad del aire sea mejorable no implica que realmente sea mejorable.

Patricia Lázaro Tello
plazarotello@uoc.edu

3 Conclusiones

- Se han cargado 7.464 registros y 23 atributos, que contenían valores nulos en varias columnas.
- Se ha estudiado las **correlaciones lineales** de las variables por registro y su media diaria. Se observa que la media diaria suaviza las correlaciones entre variables debido a que algunas mediciones con más acusadas durante las horas de luz (como la radiación solar) y otras persisten día y noche (la concentración de O_3).
- Se han realizado modelos de regresión lineal para predecir la concentración de O_3; sin embargo, **no existe una única variable predictora que pueda explicar adecuadamente la concentración de O_3**. Se encuentra además que las partículas en suspensión explicación más variación de la concentración de O_3 que la radiación solar.
- Se han creado varios **modelos de regresión lineal múltiple** para predecir la concentración de O_3. **Ningún modelo de los planteados es adecuado** por no explicar un porcentaje de variación de la concentración de O_3 aceptable, pero se ha visto que un punto de partida para la creación de un modelo de regresión lineal múltiple podría ser tomar como variable predictora la concentración de NO_2.
- El **Factor de Inflación de la Varianza** (FIV) puede no ser suficiente para decidir si introducir o no una variable al modelo ya que la relación entre las variables puede no ser lineal; es importante comprobar la idoneidad de la variable con otros tests, como la relevancia estadística del predictor al introducirlo en el modelo.
- Se ha diagnosticado el mejor modelo de regresión lineal múltiple ($modelo_3$). Los resultados del diagnóstico concluyen que **el modelo no se ajusta bien a los datos**, así como una ausencia de homocedasticidad y normalidad en sus residuos.
- Del análisis crudo de posibles factores de riesgo se obtiene que tanto la radiación solar como la velocidad del viento son **factores de riesgo**. Respecto a la humedad relativa, se puede asumir que también es factor de riesgo.
- Se han creado varios modelos de regresión logística y se ha estudiado las posibles interacciones y confusiones entre variables. Se concluye que **la temperatura es una variable de confusión** respecto a la radiación solar y que existe **interacción entre la humedad relativa y la radiación solar**. La dirección del aire mejora el modelo inicial, que utilizaba únicamente la radiación solar pare

Patricia Lázaro Tello
plazarotello@uoc.edu

predecir si la calidad del aire basado en O_3 era mejorable o buena.

- Se ha calculado la bondad del ajuste y la curva ROC del modelo final ($modelo_4$), y se ha obtenido como conclusión que **el modelo no ajusta bien a los datos** y el umbral de clasificación ha de ser muy cercano al 100% de probabilidad para conseguir un compromiso entre sensibilidad y especificidad.